The LINGBAN System for Blizzard Challenge 2019

Yansuo Yu, Fengyun Zhu

Lingban Research, Beijing, P.R. China

ysyu@ling-ban.com

Abstract

This paper introduces a text-to-speech (TTS) system developed at LINGBAN research for the Blizzard Challenge 2019. The task for this year is to build a voice from about eight hours of highly expressive Mandarin speech data. We proposed a neural vocoder based parametric system that modeling speech waveforms for this task. Firstly, A lightly-supervised speech recognition approach was adopted to select the clean speech data with accurate text. Moreover, a hybrid deep neural network (DNN) with long-short term memory (LSTM) built on multi-speaker speech data was applied for acoustic modeling and duration modeling. Finally, a WaveNet-based neural vocoder was used to generate speech waveforms from acoustic feature instead of the conventional vocoder. Subjective evaluation results show that our system performs good in all evaluation criteria.

Index Terms: speech synthesis, Blizzard Challenge, DNN-LSTM, WaveNet

1. Introduction

Speech synthesis has made great progress in the past two decades. The quality of synthesized speech has been significantly improved. Accordingly, the demand for offering high-quality synthetic speech with various speaking styles and various languages is increasing. However, due to the differences in corpus, tasks and listening test, it becomes more difficult to compare different synthetic systems. Therefore, Blizzard Challenge [1] was organized annually since 2005 to better understand and compare research techniques in constructing corpus-based speech synthesis systems with the same data. A series of tasks, such as English, Mandarin, some Indian languages, English audiobooks, have helped us measure progress in speech synthesis technology.

Early challenges tended to have a relatively small amount of speech data recorded with less noise under the same recording conditions. At that time unit selection based waveform concatenation approaches [2, 3] and hidden Markov model (HMM) based speech synthesis approaches [4, 5, 6] became the most popular methods. The biggest drawback for waveform concatenation approaches is the demand of large speech corpus with fine labeling and expert fine-tuning. HMM-based speech synthesis systems have advantages on flexibility and small footprint, but the speech quality of these systems is limited by accuracy of acoustic modeling and the traditional vocoder [6]. Afterwards in the Blizzard Challenge 2013 [7, 8], approximately 300 hours of unsegmented audio were provided as training data. The corpus of this challenge had the following characteristics, such as inconsistency between speech and text, and great change of voice expression. In the Blizzard Challenge 2016 [9], 2017 [10] and 2018 [11], similar highly expressive speech data from professionally produced English childrens audiobooks were also provided as training data. Meanwhile, the methods based on big data have achieved great success in various research fields, such as speech recognition [12] and speech synthesis [13]. The increase of speech corpus not only significantly improved the quality of synthesized speech but also made speech synthesis with various speaking styles become possible. Recent successes of deep learning methods for TTS further lead to high-fidelity speech synthesis. A variety of deep neural networks models, such as DNN-BLSTM [14] and LSTM-RNN [15], have achieved greater performance in acoustic modeling and duration modeling. In addition, the neural network based vocoders, such as WaveNet [16], SampleRNN [17], Parallel WaveNet [18] and ClariNet [19], play a very important role in recent advances of speech synthesis.

A neural vocoder based statistical parametric system has been submitted for the Blizzard Challenge 2019. Our system have three main components: data selection and text analysis, acoustic features prediction, and speech waveform generation. In the data selection and text analysis, the clean speech data with accurate text is first selected according to the word error rate (WER) of recognition results. Then linguistic contextual features, such as phonemes(initials and finals for Mandarin), tones, syllables, word segmentation, and parts-of-speech, is estimated from the input text. In the acoustic features prediction component, duration and acoustic models based on DNN-LSTM successively predict duration and acoustic features with the corresponding linguistic contextual features. In the process of speech waveform generation, a speech waveform is generated from the acoustic features by using WaveNet-based neural vocoder.

The rest of this paper is organized as follows. Section 2 introduces the details of the single Mandarin task in Blizzard 2019. An overview of our system will be discussed thoroughly in Section 3. The results of the evaluation are further described in Section 4. Finally, the conclusion is drawn in Section 5.

2. The Mandarin Task in Blizzard 2019

In Blizzard Challenge 2019, the evaluation only consists of one task as follows:

- MH1 - About 8 hours of speech data from an internet talk show by a well-known Chinese character will be released. All data are from a single speaker. The task is to build a voice from this data that is suitable for expressive TTS.

In the following sections we will introduce the whole process of constructing the speech synthesis system for MH1.

3. Overview of the System

The overview of the text-to-speech (TTS) system, which consists of both training and synthesis phases, is shown in Figure 1. At training stage, the clean speech with accurate text is firstly chosen by means of lightly-supervised speech recognition and text alignment. Afterwards the acoustic features including spectral envelope and fundamental frequency (F_0) are extracted correspondingly and the contextual labels including

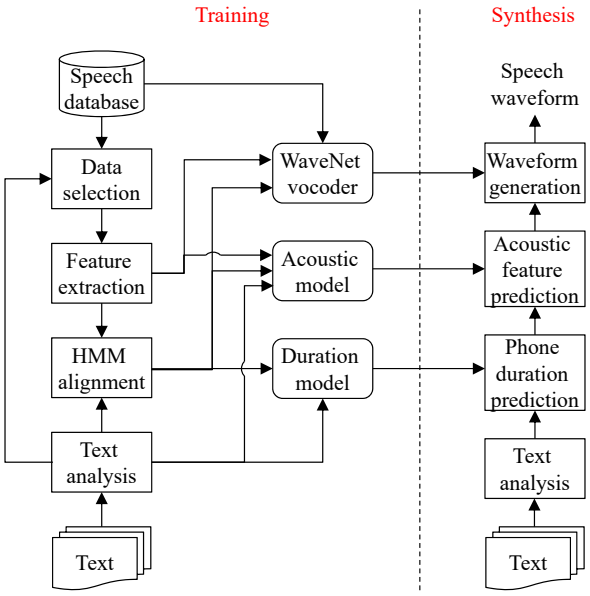

Figure 1: *The flowchart of LINGBAN TTS system.*

phone-related and word-related features are obtained through the modules of text analysis, such as text normalization, word segmentation, part-of-speech tagging, phonetic disambiguation and others. Based on these acoustic features and the contextual labels, the corresponding HMMs are estimated in the maximum likelihood (ML) sense [4] and thus obtained frame-level alignments between speech and text. Thirdly, both the duration model and acoustic model represented by the same DNN-LSTM framework are first initialized with multi-speaker speech data and further optimized using speech of target speaker. At last, a WaveNet neural vocoder conditioned on the ground-truth mel-spectrograms plus F_0 is achieved under the alignment of speech waveform and acoustic features.

At synthesis stage, the contextual label sequence of synthesized text is first predicted by the front-end text analysis. Then phone durations and acoustic features corresponding to this label sequence are successively predicted by previously trained DNN-LSTM models. Finally the WaveNet neural vocoder is used to generate speech waveforms sample by sample conditioning on the predicted mel-spectrograms plus F_0. The following subsections will introduce the whole system in detail.

3.1. Data Selection and Text Analysis

Through detailed analysis about the speech data of MH1 task, it can be found that this corpus has the following characteristics: (1) it contains a total of 480 mp3 files, each containing about one minute of speech; (2) all the speech data isn't transcribed; (3) the average gain varies from one speech segment to another due to different recording environment; (4) various voices including laughter and complex emotional expressions exist in this internet talk show. Hence it's necessary to first choose clean speech data with quite accurate transcription from raw corpus in order to construct the subsequent speech synthesis system.

Referring to [20, 21], the basic process of speech data selection is designed as follows. The untranscribed speech firstly roughly hand-annotated. The related models of speech recognition are trained based on the annotated speech. Thus if the recognition result is not identical with raw transcription, it's likely that the transcription has the errors, such as insertion error, deletion error or substitution error. Afterwards, the word error rate (WER) for each utterance is calculated through text alignment and the corresponding clean speech with accurate transcription are also obtained.

Different from previous challenges, this challenge includes a variety of synthesized texts, such as story, encyclopedias, poetry, English mixed reading, rhotic accent and others. Hence the text analysis with higher accuracy and robustness becomes particularly important. The whole process of text analysis consists of several steps. The raw texts are first converted to pure Chinese characters through the fine-designed rules. Afterwards word segmentation of the sentence, Part-of-Speeches (POS) of this word sequence and prosodic hierarchy are successively predicted by the pre-trained neural network models. In the module of grapheme to phoneme (G2P), tonal syllable sequence corresponding to the word sequence is obtained with the specific multi-pronunciation models and pronunciation dictionary. Finally, a series of contextual labels, such as positional features, numerical features and category features, are extracted from the above phonetic and linguistic information.

3.2. DNN-LSTM based Acoustic Modeling

The weakness of conventional HMM-based acoustic modeling is the accuracy of acoustic modeling, which generates the over-smoothed spectral envelopes and finally leads to the muffled voice quality of the synthetic speech. In contrast, DNN-LSTM or LSTM-RNN, which could model temporal sequences and their long-term dependencies, has been successfully applied to acoustic modeling and proven the superiority of producing more natural synthetic speech.

This work adopted same hybrid structure of the DNN-LSTM for both acoustic modeling and duration modeling. The basic DNN-LSTM, which is configured with two feed-forward layers, one unidirectional LSTM layer and one feed-forward output layer, is firstly estimated from hundreds of hours of multi-speaker speech data. Based on the pre-trained basic DNN-LSTM without output layer, the final DNN-LSTM added with one unidirectional LSTM layer and one feed-forward output layer is further trained using the speech of target speaker. For the duration modeling, both the feed-forward layer and unidirectional LSTM layer consist of 64 nodes. Previous extracted linguistic features are adopted as input feature and phone durations are used as final output. While for the the acoustic modeling, both the feed-forward layer and unidirectional LSTM layer contains 256 nodes. Except for linguistic features, duration features are also employed for input. In addition, all the above networks are trained using Stochastic Gradient Descent (SGD) algorithm. The training process will stop if no new best error on the validation set could be achieved within the last 30 epochs.

3.3. WaveNet-based Neural Vocoder

Due to lack of phase prediction and inherent assumption of vocoder, conventional vocoders like STRAIGHT [22] face great difficulties in producing high fidelity speech. To address this problem, this work adopts WaveNet generative model for waveform generation instead.

WaveNet is a fully probabilistic and autoregressive generative model that can generate waveforms directly. Given a waveform $\mathbf{x}=\{x_1, x_2, \cdots, x_T\}$, the joint probability of all these samples is factorised as a product of conditional probabilities

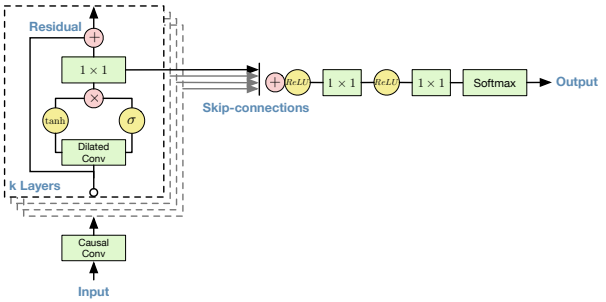

Figure 2: *Overview of the WaveNet architecture [16].*

as follows:

$$p(\mathbf{x}|\mathbf{c}) = \prod_{t=1}^{T} p(x_t|x_{<t}, \mathbf{c}; \theta) \quad (1)$$

where each audio sample x_t is conditioned on the samples at all previous timesteps. $\mathbf{c}$ is conditional inputs, here both mel-spectrograms and F_0 are adopted as local condition.

As shown in Figure 2, $p(x_t|x_{<t}, \mathbf{c}; \theta)$ in equation (1) is characterized by a stack of convolutional layers containing gated activation unit as the output of each layer:

$$\mathbf{z} = \tanh(\mathbf{W}_{f,k} * \mathbf{x} + \mathbf{V}_{f,k} * \mathbf{c}) \odot \sigma(\mathbf{W}_{g,k} * \mathbf{x} + \mathbf{V}_{g,k} * \mathbf{c}) \quad (2)$$

where $*$ denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, $\mathbf{W}_{f,k}$ and $\mathbf{W}_{g,k}$ are learnable convolution filters for waveform inputs, $\mathbf{V}_{f,k}$ and $\mathbf{V}_{g,k}$ are learnable convolution filters for conditional inputs. In addition, both residual and skip connections are utilized to further speed up convergence and enable training of much deeper models.

Finally, our WaveNet model consists of 30 layers, grouped into 3 dilated residual block stacks of 10 layers. For every stack, the dilation rate increases by a factor of 3 in every layer, and no dilation for the first layer. The number of hidden units both in the gating layers and in the residual connection is 512, and the number of hidden units is 256 for the skip connection. The network was trained for 1,000,000 steps with the ADAM optimiser [23] with a mini-batch size of 2 audio clips, each containing 6139 timesteps (roughly 383ms).

4. Results and Discussion

This section will discuss the official evaluation results of our system in Blizzard Challenge 2019 in detail. 26 systems, including 2 benchmarks and 24 submitted systems were evaluated. Our system is identified as S, whereas system A and B are benchmark systems. System A is the natural speech and system B is the Merlin Benchmark system.

4.1. Similarity test

Figure 3 shows the boxplot results of similarity scores of all systems for MH1. It can be seen that our system achieves the second best similarity to original speaker for MH1 except for system M. Moreover the results of Wilcoxons signed rank tests further show that the difference between system S and system M on similarity is not significant for MH1. Besides, system I and system W, which performed good for MOS on naturalness, instead achieve relatively poor performance.

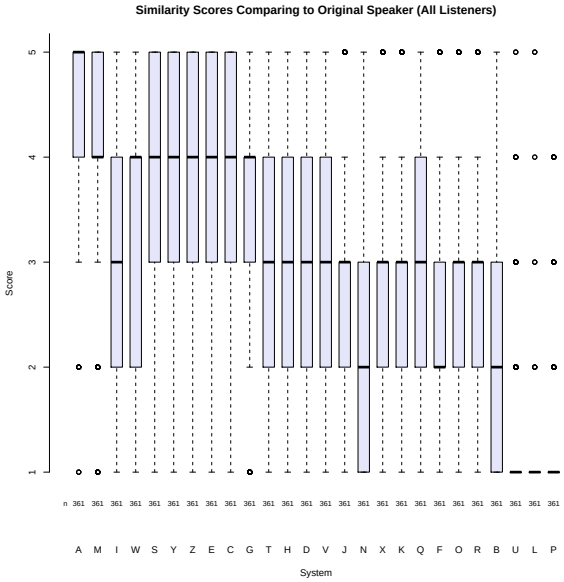

Figure 3: *Results of MOS on speaker similarity for MH1.*

4.2. Naturalness test

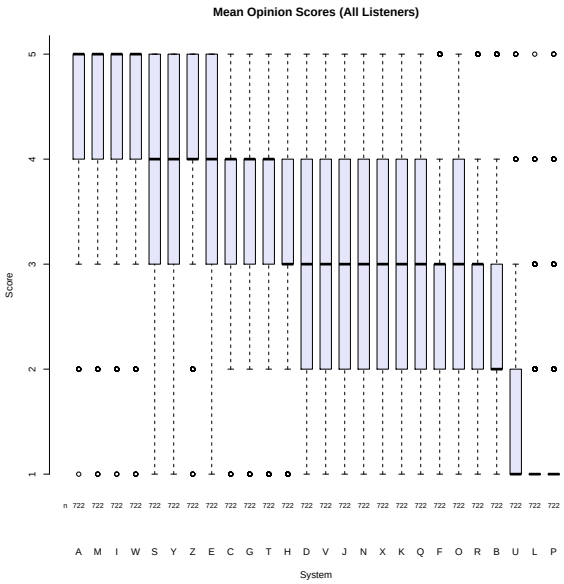

Figure 4: *Results of MOS on naturalness of sentences for MH1.*

Figure 4 shows the boxplot results of MOS on naturalness of all systems for MH1. As we can see, our system achieved better performance (not including the natural speech system A) on naturalness than merling benchmark system and most participates, except for system M, I and W.

4.3. Intelligibility test

Figure 5 and Figure 6 show the results of the overall Pinyin Error Rate (PER) test and Pinyin+Tone Error Rate (PTER) test of all systems for MH1 respectively. Semantically Unpredictable

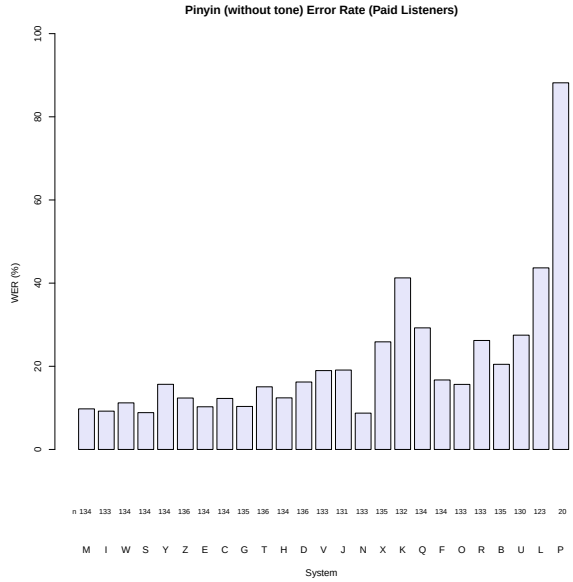

Figure 5: *Results of Pinyin Error Rate (PER) for MH1.*

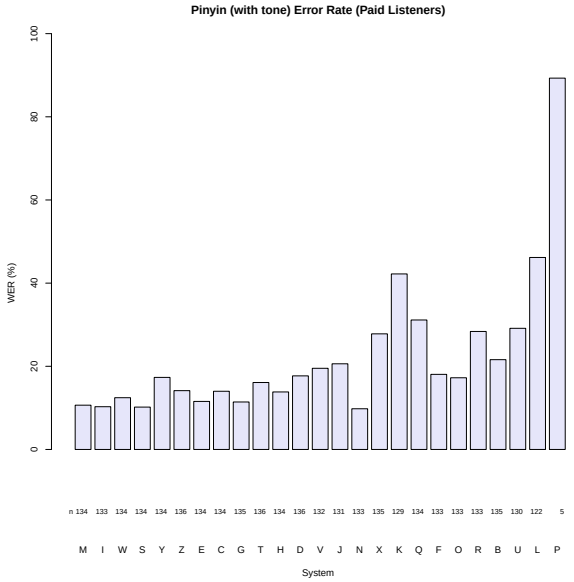

Figure 6: *Results of Pinyin+Tone Error Rate (PTER) for MH1.*

Sentences (SUS) were designed to test the intelligibility of the synthetic speech. The results show that our system achieves the second highest intelligibility among all the systems for MH1 except for system N.

5. Conclusions

This paper introduces the development of the LINGBAN speech synthesis system for Blizzard Challenge 2019. We built a neural vocoder based statistical parametric system. A lightly-supervised speech recognition approach was utilized to remove poor quality speech data with noise text. The hybrid DNN-LSTM models built on mult-speaker speech data were used for acoustic modeling and duration modeling. Due to lack of phase prediction and inherent assumptions of traditional vocoder, a WaveNet based neural vocoder was adopted to generate speech waveforms from acoustic features. The official evaluation results of Blizzard Challenge 2019 further reveal the superiority of our system. Finally our system achieves overall good performance compared to other systems according to all evaluation criteria.

6. Acknowledgements

We would like to thank Yi Liu for front-end text processing and Mandarin native speakers Ran Li, et al. for internal subjective listening test and assist in pronunciation. Thank colleagues Fan Liu and Yang Li for supporting data processing and labeling.

7. References

[1] A. W. Black and K. Tokuda, "The blizzard challenge - 2005: Evaluating corpus-based speech synthesis on common datasets,," in *INTERSPEECH*, 2005, pp. 77–80.

[2] R. E. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Cambridge University, 1996.

[3] Z. H. Ling and R. H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," in *ICASSP*, 2007, pp. 1245–1248.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, 1999, pp. 2347–2350.

[5] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *ICASSP*, 1999, pp. 229–232.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[7] S. King and V. Karaiskos, "The blizzard challenge 2013," in *Blizzard Challenge 2013 Workshop*, 2013.

[8] Y. S. Yu, F. Y. Zhu, X. G. Li, Y. Liu, J. Zou, Y. N. Yang, G. L. Yang, Z. Y. Fan, and X. H. Wu, "Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013," in *Blizzard Challenge 2013 Workshop*, 2013.

[9] S. King and V. Karaiskos, "The blizzard challenge 2016," in *Blizzard Challenge 2016 Workshop*, 2016.

[10] S. King, L. Wihlborg, and W. Guo, "The blizzard challenge 2017," in *Blizzard Challenge 2017 Workshop*, 2017.

[11] S. King, C. Jane, M. Amy, and W. Lovisa, "The blizzard challenge 2018," in *Blizzard Challenge 2018 Workshop*, 2018.

[12] E. Battenberg, J. T. Chen, R. Child, A. Coates, Y. Gaur, L. Yi, H. R. Liu, S. Satheesh, D. Seetapun, and A. Sriram, "Exploring neural transducers for End-to-End speech recognition," arXiv preprint arXiv:1707.07413, 2017.

[13] Y. S. Yu, F. Y. Zhu, X. G. Li, Y. Liu, and X. H. Wu, "Research on speech synthesis for large-scale corpora," *Journal of Natural Science of Peking University*, vol. 50, 2014.

[14] Y. C. Fan, Y. Qin, F. L. Xie, and F. K. Soong, "TTS synthesis with Bidirectional LSTM based recurrent neural networks," in *INTERSPEECH*, 2014.

[15] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *ICASSP*, 2014.

[16] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[17] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *ICLR*, 2017.

[18] A. V. D. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. V. D. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, and et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *ICML*, 2018.

[19] W. Ping, K. N. Peng, and J. T. Chen, "ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech," arXiv preprint arXiv: 1807.07281, 2018.

[20] X. G. Li, Z. H. Pang, and X. H. Wu, "Lightly supervised acoustic model training for mandarin continuous speech recognition," *Lecture Notes in Computer Science*, vol. 7751, pp. 727–734, 2013.

[21] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[22] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.